基于被引 - 逆文档权重的专家专长识别与分析*
——以图情领域为例

■ 唐晓波[1,2] 周禾深[1] 李诗轩[3] 牟昊[4]
[1] 武汉大学信息管理学院 武汉 430072 [2]武汉大学信息系统研究中心 武汉 430072
[3] 武汉理工大学安全科学与应急管理学院 武汉 430070 [4]国网四川省电力公司 成都 610000

摘 要: [目的/意义]识别专家专长有助于发现具有相同或相近研究方向的研究者,对开展细粒度的专家评价与分析具有重要意义。[方法/过程]基于学术论文关键词构建专长种子词典,采用语义相似度计算对词典进行扩展与对齐;融合专长术语被引频次、作者贡献率与专长术语逆文档频率,提出专家专长术语的被引 - 逆文档权重计算方法;结合专长权重得分及排名,识别专家的代表性研究专长,并进行专家评价与分析。[结果/结论]经实验验证,本研究提出的专家专长识别方法能够客观地反映专家专长的影响力,同时在细粒度专家评估、专家推荐以及学科热点分析等相关领域具有一定的实践参考价值。

关键词: 信息计量 语义挖掘 专长识别 专家评价
分类号: G250
DOI: 10.13266/j.issn.0252 - 3116.2021.15.013

1 引言

2020 年 10 月,中共中央国务院出台《深化新时代教育评价改革总体方案》,强调高校教师科研评价的重要性,并提出要根据不同学科、不同岗位特点,坚持分类评价,推行代表性成果评价,探索长周期评价,完善同行专家评议机制,注重个人评价与团队评价相结合[1]。然而,随着新兴学科、交叉学科的不断涌现,多样化的信息资源与科研成果数量大、种类多、更新快等特点,使得传统信息计量学方法无法满足新时代的科技人才评价需求。因此,如何应对融合态的哲学社会科学发展趋势,制定细粒度的科学人才评价管理体系,提升多元化的科技人才队伍建设水平,进而优化学科资源的利用效果,促进新时代学术科研创新发展,成为了当前亟待解决的问题。

随着"小同行"概念的提出,研究者开始对学科领域内相同或相近研究方向的专家开展细粒度评价与分析。识别专家专长可以发现"小同行"专家群体,并支持专家遴选、开展多维专家评价与分析工作。统计法是最常见的专家专长识别方法,李刚等基于词频提取专家专长,并对我国图书情报与档案管理领域的相似研究专家进行聚类及可视化分析[2];在考虑文档位置的基础上,唐晓波等通过统计医生学术成果的关键词,构建医生画像的成果特征[3];刘晓豫等将关键词作为候选专长术语,抽取作者 - 关键词矩阵,并结合 TF-IDF 加权构建专家专长[4]。部分研究者基于网络分析方法进行专长识别,朱伟珠等在词频分析的基础上构建了概念知识网络,并采用 K-core 层次理论划分学科领域的层次结构及其研究子群[5];刘萍与周梦欢提出了基于共词网络的专家专长识别方法[6]。陈翀等则将 TextRank 和概念链接技术相结合识别表示专家专长的候选专长术语,并基于署名位序与被引数等信息,使用层次分析法为专长术语分配权重[7]。此外,部分研究者基于主题分析识别专家专长,张晓娟等利用 PLSA 对每位专家的论文产出进行主题建模,并分析图情领域专家的研究领域[8],陈红伶等将 Word2vec 词向量模型与 LDA 主题模型相结合,构建专家特征并识别学术共同体[9]。目前专家专长识别方法较为局限,大部分

* 本文系国家自然科学基金项目"基于大数据的科教评价信息云平台构建和智能服务研究"(项目编号:19ZDA349)研究成果之一。
作者简介: 唐晓波(ORCID:0000 - 0001 - 5885 - 45090),教授,博士生导师;周禾深(ORCID:0000 - 0003 - 1133 - 2812),博士研究生,通讯作者, E-mail:zhouheshen@whu.edu.cn;李诗轩(ORCID:0000 - 0002 - 1879 - 4895),博士,讲师;牟昊(ORCID:0000 - 0002 - 1950 - 9953),高级工程师,博士研究生。
收稿日期:2021 - 01 - 31 修回日期:2021 - 05 - 12 本文起止页码:111 - 119 本文责任编辑:易飞

研究者采用统计术语词频的方法构建专家专长标签，且在术语权重的计算中引入了一定的主观因素。基于领域知识库识别专家专长需要集合专家知识进行领域本体构建，而基于主题分析等方法抽取的专家专长识别方法则又存在可解释性较差等问题。当前专家专长识别相关研究大多以专家研究成果的相关文本或网络关系来抽取代表性专长，忽略了成果对学科领域所产生的影响以及专家在成果中的贡献大小等因素。

因此，本研究提出了基于被引－逆文档权重的专家专长识别方法。将论文关键词与词向量模型相结合，自动构建专家专长术语词典。融合作者贡献率、被引频次与专长术语逆文档频率，提出专长术语权重计算方法。通过计算专家的专长权重得分进行排序，最终提取专家的代表性专长标签。本研究提出的专家专长识别方法能够结合相关领域研究者规模、专家在相关领域中的影响力等因素，客观地提取出专家的代表性专长，对专家评估、专家推荐与学科热点分析等方面而言具有重要实践意义。

2 相关研究

2.1 专家学术评价研究

学界针对专家评价开展了多方面的探索，传统研究者主要通过篇目分析法、引文分析法对科技人才进行评价[10]。较为经典的专家评价方法包含h指数[11]与p指数[12]，其主要通过一定时期内发表论文数及被引数等构建专家评价指标。同时部分研究者从论文数、署名位序及发表时间等方面优化评价指标并构建了衍生专家评价指数[13-15]。但刘中兴与杨建林指出，我国图情领域专家的个人学术评价指标使用仍处于发展阶段，学者们主要针对h类指数的指标开展研究，而对个人学术综合评价的多元指标融合途径研究较少，包括个人学术评价在内的学术评价研究仍需要进一步完善[16]。近年来，社会网络分析[17-18]、主题分析[8]与专家知识地图[19-20]等也逐渐成为了学科领域开展专家评价与分析的常见方法；此外，部分研究者还构建了专家知识图谱进行专家评估与分析，常见的专家知识图谱包括了基于合作关系的专家知识图谱、基于文档内容分析的专家知识图谱、基于链接分析的专家知识地图的图谱、综合内容分析和链接分析的专家知识图谱等[21]。

但是，目前专家的细粒度评价与分析研究还相对较少，由于学科或研究方向存在差异性等因素，仅以分数来评价专家的影响力是存在局限性的。同时，在专家评价相关研究中，学者通常选择特定领域的部分专家开展分析，其研究方法不能对海量专家学者进行细粒度的影响力评价。

2.2 关键词抽取与专长词典构建

基于领域知识库的专家专长表示方法能够对专家专长进行准确的描述，为构建能够反映领域知识的专长词典，需要从研究成果中抽取出能够反映和区分研究主题的术语。常见的专长词典构建方法是利用作者给出的论文关键词，如范晓玉等采用科研人员发表的文献关键词，构建专家的研究主题及兴趣标签[22]。部分研究者通过统计从论文摘要中挖掘的关键词构建专长词典，如毛进等选择专家研究成果中的高频名词代表专家的研究专长[23]。同时，陈翀等则将词共现网络与TextRank相结合来形成学术专长候选词[7]。随着自然语言处理领域的发展，一部分研究者对于如何从学术论文摘要及正文中识别关键词开展了研究，并将词向量模型[24]与深度学习模型[25]引入论文关键词抽取任务中。此外，领域知识库也受到了学者的关注，陆伟等将中国图书馆分类法与管理科学主题词表相结合，构建图情领域专家专长词典，将不同专家的研究成果进行映射[26]；胡月红和刘萍通过抽取学术论文领域术语，并基于关联规则、形式概念分析，挖掘术语间的关系，构建情报学领域本体[27]。

基于专家知识与领域本体构建术语词典的方法，不仅需要海量的专家领域知识，同时在应对新兴研究热点时往往会有迟滞性。而通过TextRank等算法或自然语言处理方法自动构建术语词典，虽然能够减少专长本体的人工标注成本，但也带来了可解释性较低、不能有效表示词与词之间关系等问题。

2.3 署名位序与作者贡献研究

在学科融合、学科交叉的背景下，越来越多的专家倾向于采用合作的方式开展研究，不同的署名位序能够直接体现专家的贡献大小。如图1所示，本研究对图书馆、情报与文献学领域发表的5万余篇论文的作者进行统计分析后发现，独立作者发表的文章数量呈递减的趋势。

署名位序往往和专家在研究中的贡献大小相关[28]，也带来了科研成果的专家贡献比例分配问题[29]。丁敬达等基于其构建的作者贡献率测度方法，提出通过计算专家按研究贡献率得分的总被引频次[30]，评价该专家在该领域的学术影响力。本研究采用N. T. Hagen提出的作者贡献率等级分配公式[31]计算专家在论文中的贡献度，将专家署名位序及贡献率

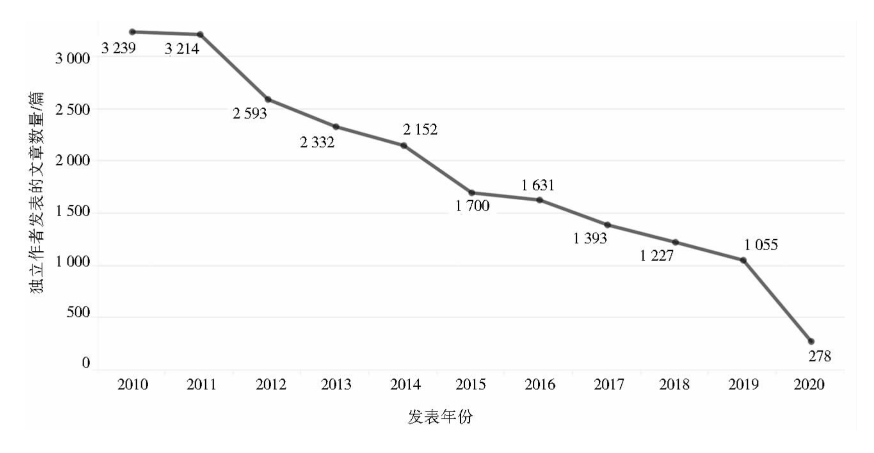


图 1　图书馆、情报与文献学独立作者发表论文统计

计算方法应用于专家专长词权重计算，从而将表示论文影响力的被引频次按照贡献率进行分配，凸显文章的重要贡献者，体现重要作者在该研究领域的科研影响力。如公式(1)所示：

$$D_j = \frac{1}{j * \sum_{j=1}^{j=m} \frac{1}{j}} \qquad \text{式(1)}$$

其中，j 代表作者的署名顺序，m 代表论文的作者总数。

3　基于被引－逆文档权重的专家专长识别模型框架

从研究成果中提取专家被研究领域所认可的研究专长是开展细粒度专家评价与分析工作的前提，本文通过对海量论文数据进行分析，将专长术语被引频次、作者贡献率与专长术语逆文档频率相结合，构建基于被引－逆文档权重的专家专长识别模型，如图 2 所示。该框架主要包括数据预处理、专长术语词典构建以及专家专长表示 3 个部分。

3.1　数据预处理

为保证数据的完备性，在数据预处理阶段将采集自多平台的中文期刊论文数据进行整合，并提取规范的学术论文数据以开展进一步分析。本文的数据预处理流程主要包括：

(1) 数据获取。基于知网、万方数据库导出目标期刊论文的元数据，采用 selenium 构建爬虫，爬取论文被引数据。

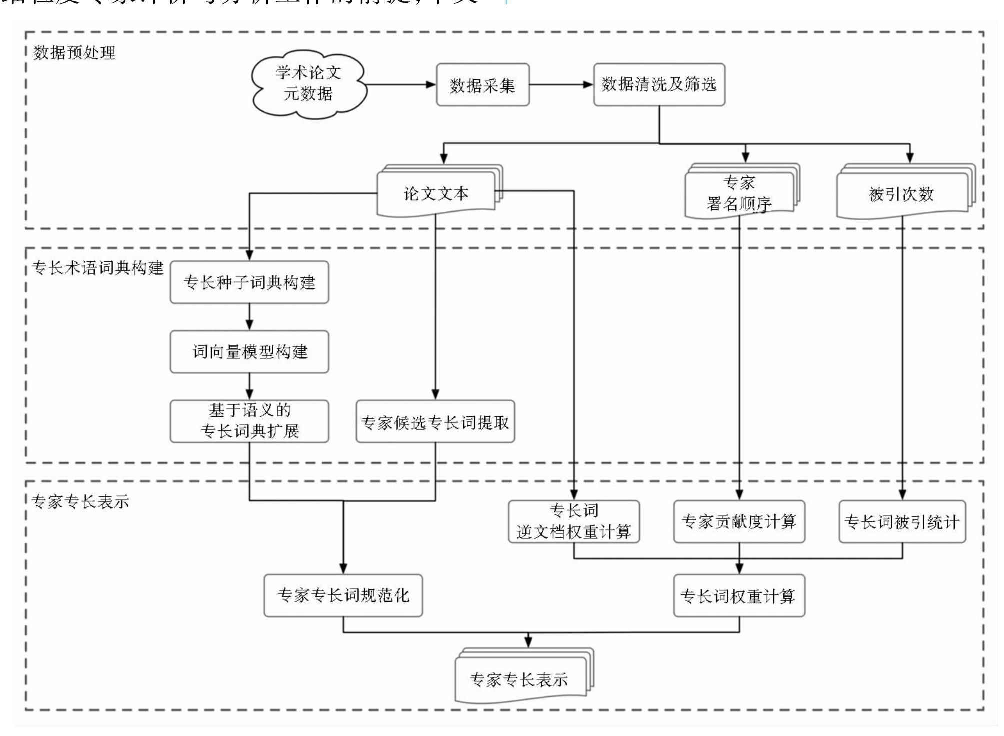


图 2　基于被引－逆文档权重的专家专长识别模型

（2）数据清洗及筛选。数据清洗主要将不同数据库论文数据进行规范化，合并数据后筛选过滤标题摘要过短、作者字段为空以及重复的样本，并定义规则去除其中的通知、收稿资讯等相关记录。

3.2 专长术语词典构建

关键词是对论文内容进行高度凝练和概括的词[22]，能够较好地反映专家的研究方向及研究能力。本研究采用领域近10年发表论文数据为研究对象，以文本中的关键词构建专长种子词典，将种子词典作为外部词典引入分词工具中，经过对摘要和标题进行分词、去停用词等预处理，构建Word2vec词向量模型。从论文标题、摘要中提取高频词作为扩展候选词，基于词向量模型进行语义相似度比较。采用与种子词典中具有高相似度的扩展候选词，建立关键词－扩展候选词同义词表。在后续的自然语言处理过程中，以同义词表将文本中异形同义的扩展候选词转化为规范化的关键词。同时，采用与种子词典中关键词相似程度均较低的候选词构建专长扩展词典，识别出与种子词典中关键词含义均不同的高频词，并通过人工过滤该词典中不能有效反映专家研究方向及研究能力的词。最后，将种子词典与专长扩展词典进行整合，得到基于语义扩展的专长词典。

3.3 专家专长表示

专家专长表示包括专长词提取及专长词权重计算两部分。在专长词提取部分，采用同义词表将原文中的高频词转化为标准化表达的专长术语，并将专长种子词典与专长扩展词典相融合，然后利用该词典标注论文数据集中的专长术语。最后，提取出各专家的的专长词与相关论文信息。

在专长权重计算部分，本研究采用专长术语所在论文被引数作为主要因素之一，以专家在相关领域所产生的影响力大小客观衡量专长的权重得分。由于词向量模型的训练语料规模存在局限性，部分在语义扩展阶段引入的词汇不能有效反映专家专长，与此同时，逆文档频率能够反映字词是否有较好的类别区分能力[32]，因此本文将逆文档频率引入专长词权重，如公式(2)所示。通过计算专长词在论文数据集中的逆文档频率，一方面能够过滤不能表征论文研究内容的常用词，另一方面能够将相关研究领域的规模作为考量因素，避免领域专家研究内容的趋同性，从而促进多研究方向共同发展。此外，在权重得分计算中引入基于专家署名位序的作者贡献率因素，能够有效凸显相关领域的重要研究者。综上，本研究提出专家专长词权重得分计算方法如公式(3)所示。选择研究领域内人数大于10人的专长词，并按照专长词权重得分进行排序，最终获得专家的代表性专长及权重得分。

$$IDF_w = \log \frac{M}{m_w} \qquad 式(2)$$

$$Score_{jw} = \sum_{i=1}^{n} (D_{ij} * cite_{ji} * IDF_w) \qquad 式(3)$$

其中，M表示全部论文总数，w表示专长术语，m_w表示包含w的论文数量，IDF_w表示专长术语w的论文逆文档频率。i表示专家n篇论文中的第i篇论文，j表示第j位专家。$Score_{jw}$表示专家j在专长术语w上的权重得分，D_{ij}表示j专家在第i篇论文中的贡献度，$cite_{ji}$表示专家j的第i篇论文的被引次数。

4 实验与结果分析

4.1 数据采集

本研究以南大核心CSSCI来源中文期刊目录（2019－2020）中图书馆、情报与文献学领域的20个期刊为研究对象，通过知网采集学术论文元数据，同时以万方进行数据补充。采集2010年1月1日至2020年4月25日期间发表的论文相关信息共54 698篇。采集字段包括来源库、题名、作者、单位、文献来源、关键词、摘要、发表时间、第一责任人、基金、年、卷、期、页码、分类号以及被引次数，元数据主要通过知网及万方提供的数据服务导出，被引次数基于Selenium构建爬虫采集。在数据预处理阶段，将万方、CNKI数据库来源的论文数据进行整合，去除标题摘要过短、作者字段为空的样本，并去除其中的通知、收稿资讯等相关记录，合并重复记录，最终获得文章共49 399篇。

4.2 实验过程

为挖掘能够描述专家专长的术语，本文以论文数据集中词频大于3的关键词构建专长种子词典，共计7 990个词。将专长种子词典导入jieba分词工具的外部词典，经对论文数据集的标题和摘要进行分词、去停用词等预处理，设定参数维度为100，上下文窗口大小为5，最低词频3次训练Word2Vec词向量模型。从标题与摘要中提取词频大于100的高频词作为扩展候选词，基于词向量模型对高频词与专长种子词典中的关键词进行语义相似度比较。若高频词能够从专长种子词典中发现相似度大于0.9的关键词，则选择最相似的关键词构建关键词－扩展候选词同义词表（见表1），共建立关键词－扩展候选词映射关系94对。若高频词与专长种子词典中的关键词相似度均低于0.6，则将该高频词纳入专长扩展词典，删除没有意义的词

唐晓波，周禾深，李诗轩，等. 基于被引－逆文档权重的专家专长识别与分析——以图情领域为例［J］. 图书情报工作，2021,65(15):111－119.

如“在内”“两种”等，最终构建包含 37 个词的专长扩展词典如“核心”“背景”和“新颖”等。最终，通过关键词－扩展候选词同义词表将论文中的高频词进行规范性表达，同时融合专长种子词典与专长扩展词典，构建基于语义扩展的专长词典，词典共包含 8 027 个词。

表 1　关键词－扩展候选词同义词表（部分）

关键词	扩展候选词
查准率	准确率
查准率	准确度
非物质文化遗产	非遗
调查问卷	问卷
相互作用	相互影响

首先，将论文的标题与摘要进行分词、去停用词处理，其次，通过关键词－扩展候选词同义词表将其中部分的高频词替换为标准化表达的关键词，并将处理后的标题、摘要与文章的关键词进行拼接，构建该论文的词表。通过基于语义扩展的专长词典保留论文文本中选择能够较好反映专家专长的词。最后，在经过预处理的论文数据集中计算专长术语的逆文档频率。同时，提取各专家相关的署名序位、论文被引次数等信息，并基于专家署名位序计算专家在论文中的贡献率。采用公式 3 计算专家专长术语权重得分，将专家专长按照权重得分进行排序，得到专家的代表性研究专长。

4.3　结果分析

为验证本研究提出的基于被引－逆文档权重的专家专长识别方法的有效性，本文进行三部分的实证分析：首先，对本研究提出的识别方法和 TF-IDF 方法的专家专长识别效果进行对比；其次，抽取多位专家的代表性专长，并开展特定研究专长的权威研究者分析以及针对不同研究阶段专家学者的专长影响力评价；最后，选取研究领域高 h 指数专家抽取其代表性专长，进行科研团队热门研究主题分析。

4.3.1　专家专长识别对比分析

TF-IDF 算法是较为常用的专家专长识别方法之一，分为词频与逆文档频率两部分，该算法考虑了关键词对文档的重要性及类别区分能力。本研究利用 TF-IDF 方法与本研究提出方法进行专家专长识别效果的对比。将每一位专家相关的论文信息进行整合，利用基于语义扩展的专长词典构建专家关键词的 TF-IDF 矩阵。以邱均平为例，两种方法提取出的权重得分前 10 的专家专长对比如表 2 所示：

表 2　专家专长识别方法对比

专家	基于被引－逆文档权重的专家专长识别	基于 TF-IDF 的专家专长识别
邱均平	（CiteSpaceII，93.27），（高影响力作者，91.89），（学科知识扩散，67.6），（作者关键词耦合分析，63.15），（作者关键词耦合，57.35），（作者共被引分析，51.64），（替代计量学，46.61），（突变检测，40.06），（聚合模式，37.18），（计量学，34.78）	（五计学，12.57），（网络流量，8.07），（作者耦合，4.61），（资源本体，4.54），（作者关键词耦合，4.51），（替代计量，4.34），（知识交流模式，4.15），（替代计量学，4.08），（企业内部知识共享，3.99），（作者互引，3.95）

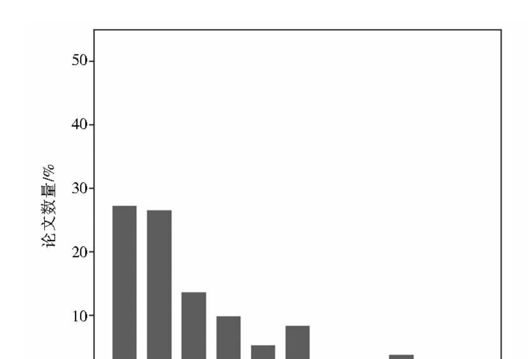
其中，基于被引－逆文档权重方法识别结果显示，邱均平在计量分析可视化和计量工具研究（CiteSpaceII）专长方面的得分最高，而基于 TF-IDF 的实验结果得出“五计学”是其具有代表性的研究专长。通过分析相关研究成果可知，邱均平在“五计学”相关领域共发表 4 篇论文，主要集中于 2019 年，且该概念的相关研究专家仅有 18 人。而基于被引－逆文档权重的方法选取了邱均平专家高被引的研究成果构建其代表性专长标签，并综合了不同专长词研究者规模因素选择专长术语，如在“CiteSpaceII”的相关研究内容中，最高被引 249 次，“学科知识扩散”相关研究分别被引 48 和 54 次。

为验证基于被引－逆文档权重方法的有效性，本研究在发文量大于 3 的专家中随机选择了 100 位专家，分别使用两种方法提取专家得分最高的专长，并对该专长的相关论文进行可视化分析，如图 3 所示。其中，被引－逆文档权重方法用以提取专家专长的论文共 132 篇，篇均被引数为 17.72 次，而 TF-IDF 方法用以提取专家专长的论文共 155 篇，篇均被引数为 8.66 次。

TF-IDF 方法用以抽取专长的论文被引数普遍较低，说明该方法在抽取专长时仅考虑了相关研究内容的数量及研究者规模，易于在研究者规模较小的研究内容中选择专长词。而本研究提出的方法所采用论文的平均被引数远高于 TF-IDF 方法。因此，本文认为基于被引－逆文档权重方法抽取的专长能够反映专家被同行所认可的代表性研究方向，并且能够挖掘出研究领域较新且认可度较高的研究主题，对于促进学科多研究方向共同发展具有重要意义。

4.3.2　专家专长评价

本文提出的专家专长识别方法，能够从多维度开展专家评价与分析。计算领域研究者的专长权重得分并排序，能够挖掘研究领域的权威专家，或评价专家在该领域的研究影响力。以“大数据”相关研究为例，将



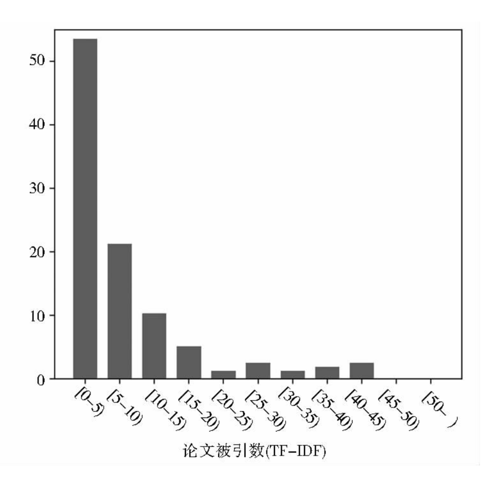


图 3　专家专长相关论文被引数分析

题名、关键词或摘要中包含“大数据”的论文作为研究对象，统计领域内的专家论文信息并计算其专长权重得分，如表 3 所示：

表 3　“大数据”相关研究专家专长权重得分

排名	专家	专长权重得分	总被引次数	相关论文篇数	篇均被引次数	一作次数
1	韩翠峰	3.455 975	433	2	217	2
2	张兴旺	2.914 427	803	19	42	13
3	苏新宁	2.746 979	415	14	30	5
4	李广建	2.701 605	611	14	41	7
5	陈臣	2.538 106	336	11	31	11

注：相关论文篇数为专家在“大数据”相关研究发表的论文数

由表 3 可得，韩翠峰仅两篇大数据研究论文，但获得了最高专长权重得分，经分析，其两篇论文分别被引 314 次和 119 次，且署名均为一作。与此同时，苏新宁虽然篇均被引数较低，但由于其在大数据研究论文中有三篇独作，最高被引 221 次，而李广建的两篇被引 178 和 165 次的一作研究论文存在共同作者，因此苏新宁在大数据领域的评分相对较高。综上可见，本研究提出的专家专长权重计算方法对高被引文章具有较强的倾向性，且对署名位序较为敏感。

此外，对专家的代表性专长及其专长权重进行分析，能够有效评价专家的学术影响力。本研究基于国内“十二五”期间 CSSCI 情报学领域高产作者与高被引作者排名、高产青年作者与高被引青年作者排名[33]，按权重得分提取专家的代表性研究专长并构建雷达图，同时展示其在该专长上的影响力排名，最终结果如图 4 和图 5 所示。通过对不同研究阶段的专家进行对比分析发现，学科高产与高被引研究专家往往在多个研究方向上均有较为深厚的学术积淀，与此同时，青年研究专家也能够通过其研究积累，在主要的部分研究方向上取得较为优秀的成绩。本研究所提出的专家专长识别方法综合考虑了专家在专长领域的贡献大小，并基于专长术语研究领域规模为专家选择了代表性专长，能够直观反映出专家研究在学科领域中的影响力，并有利于促进专家的个人成果建设，支持开展多维度的专家评价工作。

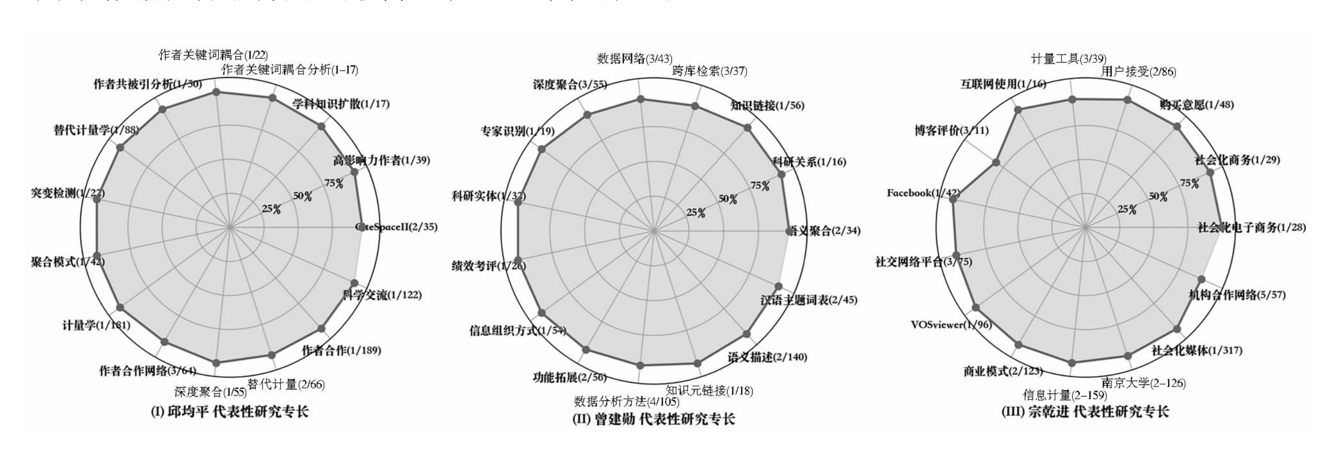


图 4　学科领域高产与高被引研究专家代表性专长雷达图(部分)

唐晓波，周禾深，李诗轩，等. 基于被引－逆文档权重的专家专长识别与分析——以图情领域为例［J］. 图书情报工作，2021,65(15):111－119.

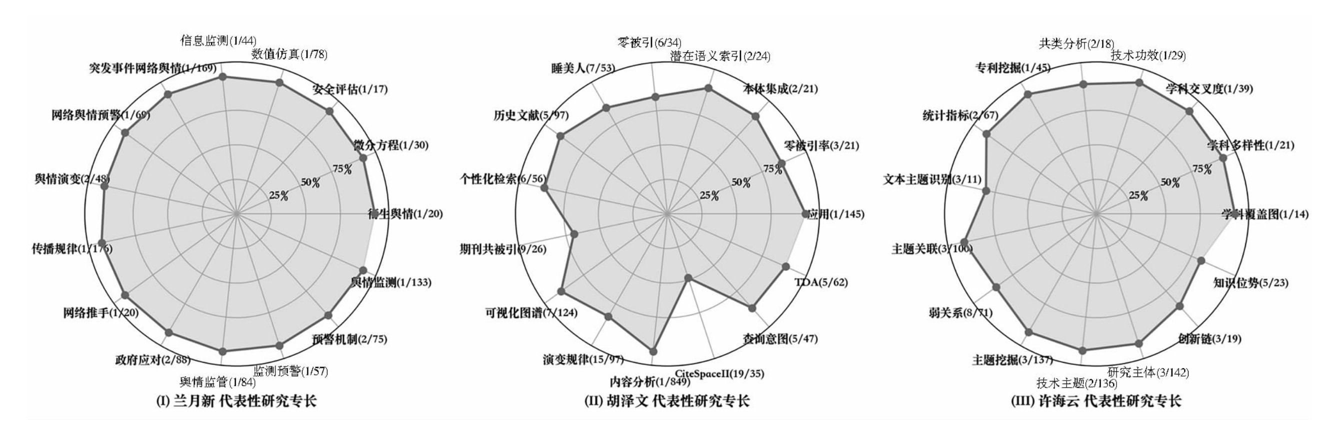


图 5　学科领域高产与高被引青年研究专家代表性专长雷达图(部分)

4.3.3　高 h 指数专家研究主题分析

引证行为在一定程度上反映了学科领域对文章内容与方向的认可，高 h 指数专家同时兼具了较高的发文量和文章被引，分析高 h 指数专家的研究内容能够有助于了解领域的热门研究。本研究以南大核心 CSSCI 来源期刊目录图书馆、情报与文献学领域在 2010 年 1 月 1 日至 2020 年 4 月 25 日期间发表的学术论文计算专家 h 指数，并对高 h 指数专家进行分析。以 h 指数得分大于等于 20 的专家为研究对象，再基于被引－逆文档权重识别上述专家的代表性专长，在表示专家专长的同义词中保留权重得分较高的专长词，最终得到领域高 h 指数专家的代表性专长及权重，如表 4 所示：

表 4　高 h 指数专家代表性专长及权重

h 指数排名	专家	h 指数得分	代表性专长及权重
1	邱均平	35	(CiteSpaceII, 93.27)，(高影响力作者，91.89)，(学科知识扩散，67.6)，(作者关键词耦合分析，63.15)，(作者关键词耦合，57.35)
2	朱庆华	28	(情感信任，29.94)，(混合方法研究，23.34)，(群体协作，21.61)，(隐私行为，18.0)，(社会计算，17.31)
3	黄如花	25	(政府数据开放共享，62.25)，(中小学生，62.17)，(政府数据开放平台，40.64)，(大规模开放在线课程，29.49)，(政府数据，27.85)
3	李纲	25	(突发公共事件，63.05)，(智库产品，54.96)，(城市应急管理，52.28)，(社会结构，39.86)，(群体行为，38.05)
5	马海群	24	(开放政策，41.67)，(CiteSpace Ⅱ，40.58)，(开放数据政策，38.35)，(高校信息公开，35.1)，(网络效应，30.8)
6	赵蓉英	23	(科学引文，204.9)，(动态网络分析，166.73)，(CiteSpaceII，117.37)，(网络计量，83.48)，(ISI，53.2)
6	赵宇翔	23	(动因研究，65.29)，(数字移民，53.08)，(UGC，48.12)，(情感信任，44.72)，(关键词分析，43.82)
6	柯平	23	(认知方式，52.54)，(图书馆战略管理，43.78)，(图书馆战略，33.08)，(成本管理，32.61)，(图书情报专业学位，28.54)
6	唐晓波	23	(细粒度情感分析，44.29)，(热点挖掘，41.55)，(产品评论挖掘，36.24)，(属性抽取，29.65)，(潜在主题，25.84)
6	初景利	23	(嵌入式学科馆员，156.39)，(新型服务能力，124.11)，(图书馆发展战略，77.98)，(智慧馆员，72.47)，(调研报告，68.19)
11	许鑫	22	(信号分析，35.14)，(专题知识库，18.58)，(政府回应，17.9)，(政务信息共享，15.93)，(DC 元数据，13.35)
11	苏新宁	22	(大情报观，91.8)，(检索技术，66.73)，(大数据思维，61.24)，(资源服务，31.61)，(数字图书馆服务，31.11)
11	王晰巍	22	(情感信任，71.83)，(雾霾，36.2)，(企业信息生态系统，24.75)，(网络团购，24.69)，(低碳技术，24.31)
11	王国华	22	(舆论反转，66.65)，(议程设置，34.97)，(辟谣，29.32)，(舆情应对，23.31)，(传统媒体，20.52)
11	张晓林	22	(科研知识，131.09)，(知识计算，128.03)，(数字学术，66.48)，(合作创新，55.71)，(研究图书馆，53.92)
11	张向先	22	(政务微信公众号，49.52)，(信息生态圈，24.11)，(情感信任，19.51)，(企业信息生态系统，18.53)
11	兰月新	22	(衍生舆情，175.2)，(微分方程，152.43)，(安全评估，127.89)，(数值仿真，89.62)，(信息监测，75.29)
18	李贺	21	(情感信任，30.13)，(模糊推理，17.24)，(Web of Science，10.43)，(社交媒体倦怠，9.84)，(隐私计算，8.58)
18	孙建军	21	(TTF，23.55)，(期望确认模型，18.51)，(用户接受模型，17.13)，(任务技术适配模型，16.31)，(期刊共被引，13.89)
18	刘炜	21	(数字对象，61.18)，(AR 技术，57.98)，(图书情报界，47.83)，(规范控制，43.48)，(语义链接，38.34)
21	马费成	20	(概念网络，32.88)，(信息老化，28.99)，(信息生命周期管理，25.58)，(演化网络，25.04)，(用户满意度模型，24.88)
21	邓胜利	20	(交互学习，31.13)，(网络社群，28.46)，(社会性网络服务，21.38)，(健康信息搜寻，20.56)，(信息源选择，19.21)
21	袁勤俭	20	(数据治理框架，16.76)，(大情报观，15.4)，(社会化电子商务，14.19)，(德尔菲法，12.63)，(南京大学，12.22)
21	肖希明	20	(公共数字文化资源，52.47)，(公共文化空间，44.06)，(LAM，31.6)，(元数据互操作，30.66)，(数字化服务，24.69)
21	王世伟	20	(复合图书馆，243.73)，(网络空间安全，182.77)，(节能，143.7)，(智能图书馆，139.15)，(智能技术，115.23)

经分析可以发现,图书馆、情报与文献学领域高 h 指数专家的主要研究领域包括了信息计量、政府数据公开、突发事件与应急响应、用户行为研究、社交媒体研究、数据分析与知识发现及图书馆管理与分析等方面。其中,信息计量、图书馆管理与分析工作获得了较高的权重得分。h 指数在专家评价工作中不能体现出专家在具体研究方向上的贡献,仍需要人工筛选评价对象与研究数据,才能够开展特定研究方向的专家评价与分析工作。本研究提出的专长识别方法是对专家评价研究的有效补充,能够从专家各研究方向所产生的影响力来丰富专家分析与评价工作。

5 结语

本文基于词向量模型构建了描述专家专长的词典,并将专长术语被引频次、作者贡献率与专长术语逆文档频率计算公式相融合,提出了基于被引 - 逆文档权重的专家专长识别方法。该方法能够基于专家的代表性研究成果提取专家专长,同时考虑研究者规模和论文影响力等因素,从学科领域影响力维度丰富了现有的专家专长识别方法。同时,该方法能够挖掘特定专长的权威专家、开展细粒度的专家评价以及分析学科领域热点等。实验结果初步验证了本研究所提出的专家专长识别方法的有效性,为专家评价与学科分析提供了新视角。

但本文所构建的专家专长识别方法仍存在一定不足,例如数据集仅采用了南大核心 CSSCI 来源中文期刊目录相关论文,且在数据选择过程中未对综述、实证研究等不同类型的论文加以区分。此外,通过论文关键词与词向量模型语义扩展构建专长词典,不能有效地揭示学科与术语的关系,不能有效区分描述研究主题和研究方法的术语,以及存在部分细粒度专长术语仍然需要专家知识进行解析才能够较好地描述专家专长。因此,如何融合学科领域知识本体,进一步优化专家专长识别方法,构建更加全面系统的细粒度专家评价模型,还有待进一步的研究探索。

参考文献:

[1] 国务院办公厅. 深化新时代教育评价改革总体方案[EB/OL]. [2021 - 01 - 15]. http://www.gov.cn/zhengce/2020-10/13/content_5551032.htm.

[2] 李刚,余益飞,杜雯. 高校 LIS 教师群体中的"小同行"研究(2001 - 2010 年)[J]. 图书情报知识,2011(6):78 - 85.

[3] 唐晓波,高和璇. 基于特征分析和标签提取的医生画像构建研究[J]. 情报科学,2020,38(5):3 - 10.

[4] 刘晓豫,朱东华,汪雪锋,等. 多专长专家识别方法研究——以大数据领域为例[J]. 图书情报工作,2018,62(3):55 - 63.

[5] 朱伟珠,李春发. 基于概念知识网络的"小同行"评议专家遴选方法实证研究[J]. 情报杂志,2017,36(7):78 - 83,88.

[6] 刘萍,周梦欢. 基于共词网络的专家专长挖掘[J]. 情报科学,2012,30(12):1815 - 1819.

[7] 陈翀,李楠,梁冰,等. 基于成果特征的学者学术专长识别方法[J]. 图书情报工作,2019,63(20):96 - 103.

[8] 张晓娟, 陆伟, 程齐凯. PLSA 在图情领域专家专长识别中的应用[J]. 现代图书情报技术, 2012(2):76 - 81.

[9] 陈红伶,杨佳颖,许鑫. 基于题录摘要语义建模的学术共同体识别——以国内图情领域学者为例[J]. 情报理论与实践,2020,43(5):170 - 176.

[10] 邱均平. 信息计量学概论[M]. 武汉:武汉大学出版社, 2019.

[11] HIRSCH J E. An index to quantify an individuals scientific research output[J]. Proceedings of the national academy of sciences, 2005, 102(46): 16569 - 16572.

[12] PRATHAP G. The 100 most prolific economists using the p-index [J]. Scientometrics, 2010, 84(1):167 - 172.

[13] YAN M, YU Z, ZHANG Y, et al. An expert recommendation approach combining project correlation and professional ability[C]// 2015 12th international conference on fuzzy systems and knowledge discovery. Zhangjiajie: IEEE, 2015.

[14] 唐璞妮. p_r(y)指数和 h_r(y)指数在学者学术影响力动态评价中的应用研究——以图情领域为例[J]. 情报理论与实践, 2020,43(12):63 - 67,41.

[15] 谢瑞霞,李秀霞,韩霞,等. 基于加权被引频次与署名顺序的作者影响力评价指标构建[J]. 情报科学,2018,36(8):90 - 93, 111.

[16] 刘中兴,杨建林. 我国图书情报领域个人学术评价指标的应用情况研究[J]. 现代情报,2020,40(12):140 - 149.

[17] 吕鹏辉, 张凌. 学科知识网络研究(Ⅱ)共被引网络的结构、特征与演化[J]. 情报学报,2014,33(4):349 - 357.

[18] 吕鹏辉, 刘盛博. 学科知识网络实证研究(Ⅳ)合作网络的结构与特征分析[J]. 情报学报,2014,33(4):367 - 374.

[19] 胡元蛟,王昊. 面向 CSSCI 的学者知识地图构建与分析[J]. 现代图书情报技术,2011(3):38 - 43.

[20] 刘勘, 周丽红. 面向专家的知识地图研究[J]. 情报资料工作, 2012, 33(2):18 - 22.

[21] 潘有能, 贺焕振. 基于合著与引用加权的专家知识地图构建研究[J]. 情报杂志, 2018, 037(8):128 - 132.

[22] 范晓玉,窦永香,赵捧未,等. 融合多源数据的科研人员画像构建方法研究[J]. 图书情报工作,2018,62(15):31 - 40.

[23] 毛进, 李纲. 一种基于 OKM 的研究领域专家图谱构建方法[J]. 图书情报工作, 2014(14):34 - 40.

[24] 徐文海,温有奎. 一种基于 TF-IDF 方法的中文关键词抽取算法[J]. 情报理论与实践,2008(02):298 - 302.

[25] ZHU X, LYU C, JI D, et al. Deep neural model with self-training for scientific keyphrase extraction[J]. Plos one, 2020, 15(5): e0232547.

［26］陆伟，刘杰，秦喜艳. 基于专长词表的图情领域专家检索与评价［J］. 中国图书馆学报，2010，36（2）：70－76.

［27］胡月红，刘萍. 基于本体概念的专长表示研究［J］. 图书情报工作，2012，56（4）：17－40.

［28］崔林蔚，陆颖. 基于作者署名排序的作者贡献要素分析——以《图书情报工作》2015－2016 年作者贡献声明为例［J］. 图书情报工作，2017，61（9）：80－86.

［29］RAHMAN M T，REGENSTEIN J M，KASSIM N L A，et al. The need to quantify authors relative intellectual contributions in a multi-author paper［J］. Journal of informetrics，2017，11（1）：275－281.

［30］丁敬达，王新明. 基于作者贡献声明的合著者贡献率测度方法［J］. 图书情报工作，2019，63（16）：95－102.

［31］HAGEN N T. Harmonic allocation of authorship credit：source-level correction of bibliometric bias assures accurate publication and citation analysis［J］. Plos one，2008，3（12）：e4021.

［32］ROBERTSON S. Understanding inverse document frequency：on theoretical arguments for IDF［J］. Journal of documentation，2004，60（5）：503－520.

［33］谭春辉，曾娟，邱均平. 基于 CSSCI 的“十二五”时期国内情报学研究态势分析［J］. 情报学报，2017，36（7）：734－747.

作者贡献说明：

唐晓波：论文总体思路、框架结构设计与修改；

周禾深：实验及论文内容组织与撰写；

李诗轩：研究设计及评测方法完善；

牟昊：论文修改及实验评测。

Identifying and Analyzing Expertise Tags of Scholars Based on the Cited-Inverse Document Frequency in the Library and Information Science Field

Tang Xiaobo[1,2] Zhou Heshen[1] Li Shixuan[3] Mou Hao[4]

[1] School of Information Management，Wuhan University，Wuhan 430072

[2] Center for Studies of Information System，Wuhan University，Wuhan 430072

[3] School of Safety Science and Emergency Management，Wuhan University of Technology，Wuhan 430072

[4] State Grid Sichuan Electric Power Company，Chengdu 610000

Abstract：［**Purpose/significance**］Identifying expertise tags helps to find scholars with the same or similar research capabilities，which is of great significance to support fine-grained scholar evaluation and analysis.［**Method/process**］In this research，we collected the keywords of academic papers to build an expertise seed dictionary，and used semantic similarity to expand and align the dictionary. Additionally，we combined the citations frequency，author contribution rate and inverse document frequency of expertise terms，and proposed cited-inverse document frequency based weight calculation method for expertise tag. Considering the weights of expertise tags，we could find the representative expertise tags of scholars，and carry out expert evaluation and analysis.［**Result/conclusion**］Experiment proves that the proposed scholar expertise identification method can objectively reflect the influence of scholar expertise，and provide a practical reference for fine-grained scholar evaluation，expert recommendation，and field hotspot analysis and other related fields.

Keywords：informetrics　semantic mining　expertise tag identification　expert evaluation